(19)日本国特許庁(JP)　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開２０００－１４８７８８
（Ｐ２０００－１４８７８８Ａ）
(43)公開日　平成12年5月30日(2000.5.30)

(51)Int.Cl.[7]		識別記号	FI			テーマコード(参考)
Ｇ０６Ｆ	17/30		Ｇ０６Ｆ	15/40	３７０　Ｂ	5B009
	17/27		Ｇ０６Ｋ	9/20	３４０	5B029
Ｇ０６Ｋ	9/20	３４０	Ｇ０６Ｆ	15/20	５５０　Ｆ	5B075
				15/401	３１０　Ａ	

審査請求　未請求　請求項の数１１　ＦＤ　（全９頁）

(21)出願番号　特願平10-328806

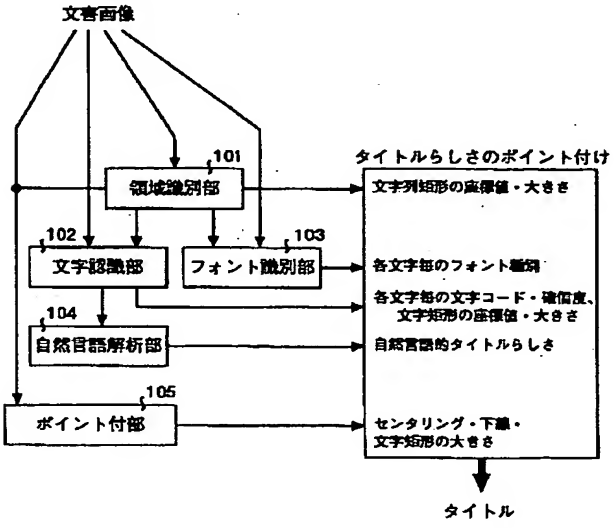
(22)出願日　平成10年11月5日(1998.11.5)

(71)出願人　000006747
株式会社リコー
東京都大田区中馬込1丁目3番6号
(72)発明者　大内　茂樹
東京都大田区中馬込1丁目3番6号　株式会社リコー内

Ｆターム(参考)　5B009 QA12
5B029 AA01 BB02 CC27
5B075 ND03 NK02 NK04 NK32 NK39
UU06

(54)【発明の名称】文書画像からのタイトル領域抽出装置およびタイトル領域抽出方法，並びに文書検索方法

(57)【要約】
【課題】　特定の文書形式に依存せずにタイトル固有の特徴をポイントとして用いることにより，ポイント数の多い文字列領域をタイトルとして自動抽出し，タイトル抽出の的確性および文書検索時の利便性を向上させること。
【解決手段】　領域識別部１０１で切り出された文字列矩形に対し，該文字列矩形内の文字認識を行う文字認識部１０２と，上記文字列矩形に対し，該文字列矩形内の各文字毎のフォント識別を行うフォント識別部１０３と，文字認識部１０２の認識結果で得られる文字コードに基づいて自然言語的タイトルらしさを解析する自然言語解析部１０４と，上記文字列矩形に対し，センタリング・下線・文字矩形の大きさ等を用いてタイトルらしさのポイント付けを行うポイント付部１０５と，を備えた。



1

【特許請求の範囲】

【請求項１】　画像入力装置から入力された文書画像から文字列領域を矩形で切り出す領域識別手段を有し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出装置において，前記領域識別手段で切り出された文字列矩形に対し，該文字列矩形内の文字認識を行う文字認識手段と，前記領域識別手段で切り出された文字列矩形に対し，該文字列矩形内の各文字毎のフォント識別を行うフォント識別手段と，前記文字認識手段の認識結果で得られる文字コードに基づいて自然言語的タイトルらしさを解析する自然言語解析手段と，前記領域識別手段で切り出された文字列矩形に対し，センタリング・下線・文字矩形の大きさ等を用いてタイトルらしさのポイント付けを行うポイント付手段と，を備えたことを特徴とする文書画像からのタイトル領域抽出装置。

【請求項２】　入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域内の文字コードを認識し，文字コード識別の確信度が一定のしきい値以上であるか否かを判断する第１の工程と，前記第１の工程で一定のしきい値以上である場合，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第２の工程と，を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項３】　入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域内の文字認識を実行し，該文字認識時に文字列矩形内の文字数を求める第１の工程と，文書のタイトルの文字数を用い，前記文字数と比較し，文字矩形数が所定値内であるか否かを判断する第２の工程と，前記第２の工程で，文字矩形数が所定値内である場合に，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項４】　入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域内の文字コードの認識結果に対して自然言語処理を実行する第１の工程と，前記第１の工程の結果，体言止めになっている領域であるかを判断する第２の工程と，前記第２の工程で体言止めになっている領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むことを特徴とする文書画像からのタイトル領域抽出方法。

2

【請求項５】　入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域内の文字コードの認識結果に対して自然言語処理を実行する第１の工程と，前記第１の工程の結果，タイトルに頻出する語尾の統計情報辞書と前記文字列領域内の文字コード列とを比較し，高頻出度の語尾と一致するものを語尾に含む文字列領域であるかを判断する第２の工程と，前記第２の工程の領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項６】　入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域に対してフォント識別処理を実行する第１の工程と，前記フォント識別処理の結果に基づいて，文字のフォントスタイルを判別し，特定のフォントを用いている文字領域であるかを判断する第２の工程と，前記第２の工程で特定のフォントを用いている文字領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項７】　入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域に対してフォント識別処理を実行する第１の工程と，前記フォント識別処理の結果に基づいて，フォントスタイル判別時に文書全体のフォントスタイルのヒストグラムを作成しておき，出現頻度の少ないフォントスタイルを用いている文字領域であるかを判断する第２の工程と，前記第２の工程で判断した文字領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項８】　入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列矩形内の各文字矩形のアスペクト比を求める第１の工程と，前記アスペクト比に基づいて倍角文字であるかを判断する第２の工程と，前記倍角文字であると判断した文字領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項９】　入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列矩形に対して文字認識処理を実行する第１の工程と，前記文字認識処理によって空白文字以外認識された各文字矩形の横幅（縦書き時は縦幅）の合計値を求める第２の工程と，前記合計値が前記文字矩形領域のほぼ半分であるかを判断する第３の工程と，前記第３の工程でほぼ半分であると判定された文字列領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第４の工程と，を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項１０】　前記タイトルらしさのポイント加算の可否判断に用いる基準値は，ユーザ単位の入力文書形式に合わせて学習して得られる最適値とし，可変・設定されることを特徴とする請求項２ないし９の何れか一つに記載の文書画像からのタイトル領域抽出方法。

【請求項１１】　文書画像を認識し，その結果に対して言語処理を行ってキーワードを抽出する第１の工程と，前記第１の工程で抽出されたキーワードと，請求項２ないし１０の何れか一つに記載の文書画像からのタイトル領域抽出方法に基づいて抽出したタイトルとを併記する第２の工程と，前記第２の工程で併記されたタイトルを用いて文書検索を実行する第３の工程と，を含むことを特徴とする文書検索方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は，ファクシミリやイメージスキャナ等の画像入力装置から入力された文書画像データのデータベースから，検索の利便性を向上させるために，文書内容を的確に表現するような文書中の領域をタイトル領域として抽出する文書画像からのタイトル領域抽出装置およびタイトル領域抽出方法，並びに文書検索方法に関する。

【０００２】

【従来の技術】従来，文書画像を検索する際には，後の検索時の利便性を図るために，画像入力装置からの文書画像の入力とは別にオペレータが手作業で，その文書の内容を的確に表現するタイトル情報やキーワード情報を抽出／作成して付加したり，定形文書に対しては，文書中の特定の位置（文字列）をタイトル・キーワードとして切り出していた。

【０００３】また，非定形文書に対してレイアウト的特徴のみを用いてタイトルを抽出する参考技術文献が，例えば，特開平９－１３４４０６号公報の『文書画像からのタイトル抽出装置および方法』，特開平５－２７４４７１号公報の『イメージ文書のタイトル領域抽出処理方法』が開示されている。

【０００４】

【発明が解決しようとする課題】しかしながら，上記に示されるような従来の技術にあっては，オペレータによるタイトル情報やキーワード情報の付加は文書量が多くなるにしたがって作業量も増加するため，作業負担の増大化を招来させてしまう。また，特定の位置の自動切り出しを行うと，定形文書のみを対象とするので，非定形文書には利用することができず，利便性に欠けるといった問題点があった。

【０００５】従来より開示されている特開平９－１３４４０６号公報・特開平５－２７４４７１号公報にあっては，レイアウト的特徴にのみ注目してタイトル抽出を行っているため，文書内容を的確に表現するタイトルの的中率が必ずしも満足できるものではなく，後の文書検索等に支障をきたす等の問題点があった。

【０００６】本発明は，上記に鑑みてなされたものであって，特定の文書形式に依存せずにタイトル固有の特徴をポイントとして用いることにより，ポイント数の多い文字列領域をタイトルとして自動抽出し，タイトル抽出の的確性および文書検索時の利便性を向上させることを目的とする。

【０００７】

【課題を解決するための手段】上記の目的を達成するために，請求項１に係る文書画像からのタイトル領域抽出装置にあっては，画像入力装置から入力された文書画像から文字列領域を矩形で切り出す領域識別手段を有し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出装置において，前記領域識別手段で切り出された文字列矩形に対し，該文字列矩形内の文字認識を行う文字認識手段と，前記領域識別手段で切り出された文字列矩形に対し，該文字列矩形内の各文字毎のフォント識別を行うフォント識別手段と，前記文字認識手段の認識結果で得られる文字コードに基づいて自然言語的タイトルらしさを解析する自然言語解析手段と，前記領域識別手段で切り出された文字列矩形に対し，センタリング・下線・文字矩形の大きさ等を用いてタイトルらしさのポイント付けを行うポイント付手段と，を備えたものである。

【０００８】また，請求項２に係る文書画像からのタイトル領域抽出方法にあっては，入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域内の文字コードを認識し，文字コード識別の確信度が一定のしきい値以上であるか否かを判断する第１の工程と，前記第１の工程で一定のしきい値以上である場合，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第２の工程と，を含むものである。

【０００９】また，請求項３に係る文書画像からのタイ

トル領域抽出方法にあっては，入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域内の文字認識を実行し，該文字認識時に文字列矩形内の文字数を求める第１の工程と，文書のタイトルの文字数を用い，前記文字数と比較し，文字矩形数が所定値内であるか否かを判断する第２の工程と，前記第２の工程で，文字矩形数が所定値内である場合に，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むものである。

【００１０】また，請求項４に係る文書画像からのタイトル領域抽出方法にあっては，入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域内の文字コードの認識結果に対して自然言語処理を実行する第１の工程と，前記第１の工程の結果，体言止めになっている領域であるかを判断する第２の工程と，前記第２の工程で体言止めになっている領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むものである。

【００１１】また，請求項５に係る文書画像からのタイトル領域抽出方法にあっては，入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域内の文字コードの認識結果に対して自然言語処理を実行する第１の工程と，前記第１の工程の結果，タイトルに頻出する語尾の統計情報辞書と前記文字列領域内の文字コード列とを比較し，高頻出度の語尾と一致するものを語尾に含む文字列領域であるかを判断する第２の工程と，前記第２の工程の領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むものである。

【００１２】また，請求項６に係る文書画像からのタイトル領域抽出方法にあっては，入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域に対してフォント識別処理を実行する第１の工程と，前記フォント識別処理の結果に基づいて，文字のフォントスタイルを判別し，特定のフォントを用いている文字領域であるかを判断する第２の工程と，前記第２の工程で特定のフォントを用いている文字領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むものである。

【００１３】また，請求項７に係る文書画像からのタイトル領域抽出方法にあっては，入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域に対してフォント識別処理を実行する第１の工程と，前記フォント識別処理の結果に基づいて，フォントスタイル判別時に文書全体のフォントスタイルのヒストグラムを作成しておき，出現頻度の少ないフォントスタイルを用いている文字領域であるかを判断する第２の工程と，前記第２の工程で判断した文字領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むものである。

【００１４】また，請求項８に係る文書画像からのタイトル領域抽出方法にあっては，入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列矩形内の各文字矩形のアスペクト比を求める第１の工程と，前記アスペクト比に基づいて倍角文字であるかを判断する第２の工程と，前記倍角文字であると判断した文字領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むものである。

【００１５】また，請求項９に係る文書画像からのタイトル領域抽出方法にあっては，入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列矩形に対して文字認識処理を実行する第１の工程と，前記文字認識処理によって空白文字以外認識された各文字矩形の横幅（縦書き時は縦幅）の合計値を求める第２の工程と，前記合計値が前記文字矩形領域のほぼ半分であるかを判断する第３の工程と，前記第３の工程でほぼ半分であると判定された文字列領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第４の工程と，を含むものである。

【００１６】また，請求項１０に係る文書画像からのタイトル領域抽出方法にあっては，前記タイトルらしさのポイント加算の可否判断に用いる基準値は，ユーザ単位の入力文書形式に合わせて学習して得られる最適値とし，可変・設定されるものである。

【００１７】また，請求項１１に係る文書検索方法にあっては，文書画像を認識し，その結果に対して言語処理を行ってキーワードを抽出する第１の工程と，前記第１の工程で抽出されたキーワードと，請求項２ないし１０の何れか一つに記載の文書画像からのタイトル領域抽出

方法に基づいて抽出したタイトルとを併記する第２の工程と，前記第２の工程で併記されたタイトルを用いて文書検索を実行する第３の工程と，を含むものである。

【００１８】

【発明の実施の形態】以下，本発明の文書画像からのタイトル領域抽出装置およびタイトル領域抽出方法，並びに文書検索方法について添付図面を参照して説明する。

【００１９】図１は，本発明の実施の形態に係るタイトル領域抽出処理を行うシステム構成を示すブロック図である。図において，１０１はファクシミリやイメージスキャナ等の画像入力装置（図示せず）から入力された文書画像から文字列領域を矩形で切り出す領域識別手段としての領域識別部，１０２は領域識別部１０１の識別結果に基づいて文字認識を行う文字認識手段としての文字認識部，１０３は領域識別部１０１の識別結果に基づいてフォント識別を行うフォント識別手段としてのフォント識別部，１０４は文字認識部１０２の認識結果で得られる文字コードに基づいて自然言語的タイトルらしさを解析する自然言語解析手段としての自然言語解析部，１０５は従来から用いられているセンタリング・下線・文字矩形の大きさ等を用いてタイトルらしさのポイント付けを行うポイント付手段としてのポイント付部である。

【００２０】図１の構成において，画像入力装置（図示せず）から文書画像が入力されると，スキュー補正等の前処理を行い，領域識別部１０１により領域識別処理を実行し，文字列矩形の座標値・大きさの情報を得る。次いで，領域識別部１０１による領域識別処理の結果を用い，文字認識部１０２による文字認識，およびフォント識別部１０３によるフォント識別を行う。

【００２１】文字認識部１０２では各文字毎の文字コード・確信度，文字矩形の座標値・大きさがタイトルらしさのポイント付けとして得られる。また，フォント識別部１０３では各文字毎のフォント種別がタイトルらしさのポイント付けとして得られる。

【００２２】また，文字認識部１０２により得られる文字コードは，自然言語解析部１０４自然言語解析ルーチンにも供給され，自然言語的タイトルらしさ，つまり，体言止めになっている領域のタイトルらしさのポイントを与える。さらに，自然言語処理において，タイトルに頻出する語尾の統計情報辞書と文字領域内の文字コード列とを比較し，高頻出度の語尾と一致するものを語尾に含む文字列領域にタイトルらしさのポイントを与える。

【００２３】また，上述の各ポイントらしさのポイントに加え，従来から用いられているセンタリング処理・下線処理・文字列矩形の大きさ等も用いてタイトルらしさの合計ポイントを計算し，タイトルを識別する。

【００２４】次に，図３～図８に示すフローチャートを参照し，本発明の一連のタイトル抽出処理方法について順に説明する。なお，このタイトル抽出処理方は，図１の構成によって複数の組み合わせあるいは単独，あるいは選択的に行ってことができる。

【００２５】図３は，実施の形態に係る第１のタイトル抽出方法を示すフローチャートであり，文字コード識別の確信度が一定のしきい値以上であった場合にタイトルらしさのポイントを加算する例について示している。まず，文書入力装置（図示せず）から文書画像を入力し（Ｓ３０１），領域識別部１０１により文字列領域を識別する（Ｓ３０２）。続いて，上記文字列領域内の文字コードを認識し，文字コード識別の確信度が一定のしきい値以上であるか否かを判断する（Ｓ３０３）。ここで，一定のしきい値以上であると判断した場合，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する（Ｓ３０４）。

【００２６】図４は，実施の形態に係る第２のタイトル抽出方法を示すフローチャートである。まず，文書入力装置（図示せず）から文書画像を入力し（Ｓ４０１），領域識別部１０１により文字列領域を識別する（Ｓ４０２）。続いて，文字認識時に文字列矩形内の文字数を求める（Ｓ４０３）。そして，文書のタイトルの文字数を用い，上記文字数と比較し（Ｓ４０４），文字矩形数が所定値内であるか否かを判断する（Ｓ４０５）。ここで，文字矩形数が所定値内であると判断すると，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する（Ｓ４０６）。

【００２７】すなわち，文字列領域内の文字コード認識時に文字列矩形内の文字数を求め，別途辞書情報として文書のタイトルの文字数の統計を用いて比較し，タイトルらしい文字数の文字列矩形に対してタイトルらしさのポイントを与える。

【００２８】図５は，実施の形態に係る第３のタイトル抽出方法を示すフローチャートである。まず，文書入力装置（図示せず）から文書画像を入力し（Ｓ５０１），領域識別部１０１により文字列領域を識別する（Ｓ５０２）。続いて，文字列領域内の文字コードの認識結果に対して自然言語処理を行い（Ｓ５０３），所定事項の領域，例えば，体言止めになっている領域か否かを判断する（Ｓ５０４）。ここで，所定事項の領域であると判断すると，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する（Ｓ５０５）。

【００２９】また，上述の言語処理において，タイトルに頻出する語尾の統計情報辞書と文字領域内の文字コード列とを比較し，高頻出度の語尾と一致するものを語尾に含む文字列領域にタイトルらしさのポイントを与えてもよい。

【００３０】図６は，実施の形態に係る第４のタイトル抽出方法を示すフローチャートである。まず，文書入力装置（図示せず）から文書画像を入力し（Ｓ６０１），領域識別部１０１により文字列領域を識別する（Ｓ６０２）。続いて，フォント識別処理を行い（Ｓ６０３），所定のフォント（フォントスタイル）を含む領域である

か否かを判断する（Ｓ６０４）。つまり，文字のフォントスタイルを判別し，特定のフォントを用いている文字領域であるか，あるいは，フォントスタイル判別時に文書全体のフォントスタイルのヒストグラムを作成しておき，出現頻度の少ないフォントスタイルを用いている文字領域であるかを判断する。そして，これらの領域であると判断した場合に，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する（Ｓ６０５）。

【００３１】図７は，実施の形態に係る第５のタイトル抽出方法を示すフローチャートである。まず，文書入力装置（図示せず）から文書画像を入力し（Ｓ７０１），領域識別部１０１により文字列領域を識別する（Ｓ７０２）。続いて，文字列矩形内の各文字矩形のアスペクト比を求め（Ｓ７０３），アスペクト比が横：縦＝２：１に近い値となる文字矩形が文字列矩形内の文字矩形数中の一定の割合以上を占めているか否かを判断する（Ｓ７０４）。ここで，一定以上の割合を占めていれば，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する（Ｓ７０５）。

【００３２】図８は，実施の形態に係る第６のタイトル抽出方法を示すフローチャートである。まず，文書入力装置（図示せず）から文書画像を入力し（Ｓ８０１），領域識別部１０１により文字列領域を識別する（Ｓ８０２）。続いて，文字認識処理を行い（Ｓ８０３），文字認識処理によって空白文字以外認識された各文字矩形の横幅（縦書き時は縦幅）の合計値を求める（Ｓ８０４）。そして，その合計値が文字矩形領域のほぼ半分であるか否かを判断する（Ｓ８０５）。ここで，合計値が文字矩形領域のほぼ半分であれば，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する（Ｓ８０６）。

【００３３】ところで，上述した実施の形態において必要となるしきい値を固定せずに，各ユーザの入力する文書の対応させて学習し，各ユーザの使用する文書形式に対して最適なしきい値を可変的に求め，初期値から変更・設定してもよい。

【００３４】さらに，上述の如く求められる一時的なポイントを基に，図２に示すように二次的な組み合わせにより，倍角文字や均等割付けの判定を行い，それらに対してタイトルらしさのポイントを与えることも可能である。

【００３５】これを付言すると，文字コードの認識時に得られる文字列矩形内の各文字矩形領域の大きさを用い，文字矩形領域のアスペクト比を算出することによって倍角文字を判定し，該倍角文字を用いている文字列領域に対してタイトルらしさのポイントを与える。

【００３６】また，文字矩形領域とそれが属する文字列領域の大きさと，文字コードの認識時に得られる文字列矩形内の文字数とを用いて矩形内の文字密度を算出し，その値によって均等割付け判定を行う。そして，均等割付けされたと判定された文字列領域に対してタイトルらしさのポイントを与える。

【００３７】ところで，上述したタイトル領域抽出方法を用いて情報検索を行うことも実現可能である。図９は，実施の形態に係る情報検索方法を示すフローチャートである。まず，文書画像を認識し（Ｓ９０１），その結果に対して言語処理を行ってキーワードを抽出する（ＳＳ９０２）。さらに，上記抽出されたキーワードと，前述のタイトル領域抽出方法によって抽出したタイトルとを併記し（Ｓ９０３），その併記タイトルを用いて文書検索を実行する（Ｓ９０４）。これにより，検索時における利便性が向上する。

【００３８】

【発明の効果】以上説明したように，本発明に係る文書画像からのタイトル領域抽出装置（請求項１）によれば，入力された文書画像から文字列領域を矩形で切り出す領域識別手段を有し，その文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する際に，領域識別手段で切り出された文字列矩形に対し，該文字列矩形内の文字認識を行う文字認識手段と，領域識別手段で切り出された文字列矩形に対し，該文字列矩形内の各文字毎のフォント識別を行うフォント識別手段と，文字認識手段の認識結果で得られる文字コードに基づいて自然言語的タイトルらしさを解析する自然言語解析手段と，領域識別手段で切り出された文字列矩形に対し，センタリング・下線・文字矩形の大きさ等を用いてタイトルらしさのポイント付けを行うポイント付手段とを設け，特定の文書形式に依存せずにタイトル固有の特徴をポイント付けとして用いるため，ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し，かつタイトル抽出の的確性および文書検索時の利便性を向上させる装置を提供することができる。

【００３９】また，本発明に係る文書画像からのタイトル領域抽出方法（請求項２）によれば，文字列領域内の文字コードを認識し，文字コード識別の確信度が一定のしきい値以上であるかを判断し，一定のしきい値以上である場合，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出するため，ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し，かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【００４０】また，本発明に係る文書画像からのタイトル領域抽出方法（請求項３）によれば，文字列領域内の文字認識を実行し，該文字認識時に文字列矩形内の文字数を求め，文書のタイトルの文字数を用い，上記文字数と比較し，文字矩形数が所定値内であるかを判断し，文字矩形数が所定値内である場合，該当する文字列領域にタイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出するため，ポイント数の多い

文字列領域をタイトルとして自動抽出することが実現し，かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【００４１】また，本発明に係る文書画像からのタイトル領域抽出方法（請求項４）によれば，文字列領域内の文字コードの認識結果に対して自然言語処理を実行し，そのの結果，体言止めになっている領域であるかを判断し，体言止めになっている領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出するため，ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し，かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【００４２】また，本発明に係る文書画像からのタイトル領域抽出方法（請求項５）によれば，文字列領域内の文字コードの認識結果に対して自然言語処理を実行し，その結果，タイトルに頻出する語尾の統計情報辞書と文字列領域内の文字コード列とを比較し，高頻出度の語尾と一致するものを語尾に含む文字列領域であるかを判断し，その領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出するため，ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し，かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【００４３】また，本発明に係る文書画像からのタイトル領域抽出方法（請求項６）によれば，文字列領域に対してフォント識別処理を実行し，その結果に基づいて，文字のフォントスタイルを判別し，特定のフォントを用いている文字領域であるかを判断し，特定のフォントを用いている文字領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出するため，ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し，かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【００４４】また，本発明に係る文書画像からのタイトル領域抽出方法（請求項７）によれば，文字列領域に対してフォント識別処理を実行し，その結果に基づいて，フォントスタイル判別時に文書全体のフォントスタイルのヒストグラムを作成しておき，出現頻度の少ないフォントスタイルを用いている文字領域であるかを判断し，該判断した文字領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出するため，ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し，かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【００４５】また，本発明に係る文書画像からのタイトル領域抽出方法（請求項８）によれば，文字列矩形内の各文字矩形のアスペクト比を求め，そのアスペクト比に基づいて倍角文字であるかを判断し，倍角文字であると判断した文字領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出するため，ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し，かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【００４６】また，本発明に係る文書画像からのタイトル領域抽出方法（請求項９）によれば，文字列矩形に対して文字認識処理を実行し，文字認識処理によって空白文字以外認識された各文字矩形の横幅（縦書き時は縦幅）の合計値を求め，その合計値が文字矩形領域のほぼ半分であるかを判断し，ほぼ半分であると判定された文字列領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出するため，ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し，かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【００４７】また，本発明に係る書画像からのタイトル領域抽出方法（請求項１０）によれば，請求項２ないし９の何れか一つに記載の文書画像からのタイトル領域抽出方法において，タイトルらしさのポイント加算の可否判断に用いる基準値を，ユーザ単位の入力文書形式に合わせて学習して得られる最適値を用いて可変・設定することにより，より的確なタイトルの自動抽出が実現する。

【００４８】また，本発明に係る文書検索方法（請求項１１）によれば，文書画像を文字認識し，その結果に対して言語処理を行って抽出されたキーワードと，請求項２ないし９の何れか一つに記載の文書画像からのタイトル領域抽出方法に基づいて抽出したタイトルとを併記し，該併記されたタイトル，すなわち，より的確なタイトルを用いて文書検索を実行するため，文書検索時における利便性が向上する。

【図面の簡単な説明】

【図１】本発明の実施の形態に係るタイトル領域抽出処理を行うシステム構成を示すブロック図である。

【図２】本発明の実施の形態に係るタイトル領域抽出処理に用いられるタイトルらしさのポイントうち，二次的に求められるものを示すブロック図である。

【図３】本発明の実施の形態に係る第１のタイトル抽出方法を示すフローチャートである。

【図４】本発明の実施の形態に係る第２のタイトル抽出方法を示すフローチャートである。

【図５】本発明の実施の形態に係る第３のタイトル抽出方法を示すフローチャートである。

【図６】本発明の実施の形態に係る第４のタイトル抽出方法を示すフローチャートである。

【図７】本発明の実施の形態に係る第５のタイトル抽出方法を示すフローチャートである。

【図８】本発明の実施の形態に係る第６のタイトル抽出

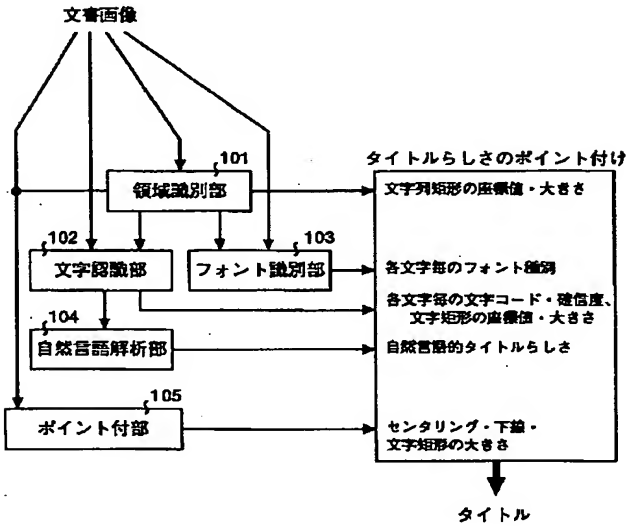
方法を示すフローチャートである。

【図９】本発明の実施の形態に係る情報検索方法を示すフローチャートである。

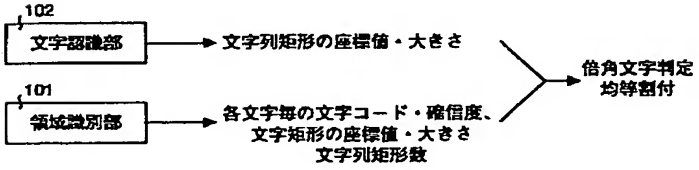
【符号の説明】

１０１　領域識別部
１０２　文字認識部
１０３　フォント識別部
１０４　自然言語解析部
１０５　ポイント付部

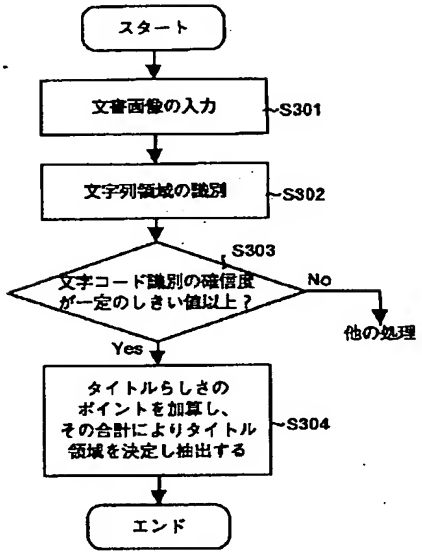
【図１】



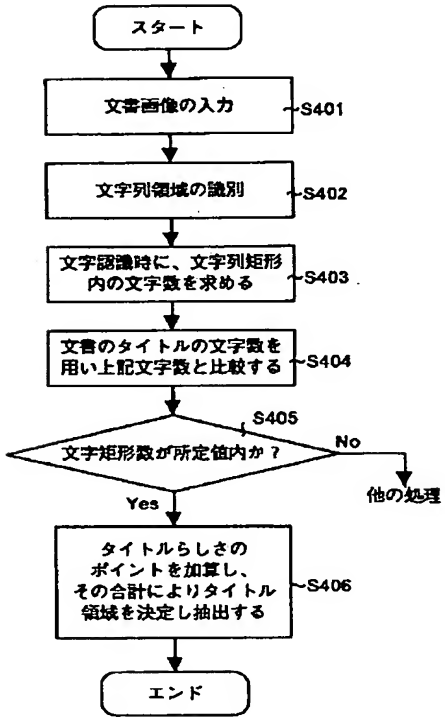
【図２】



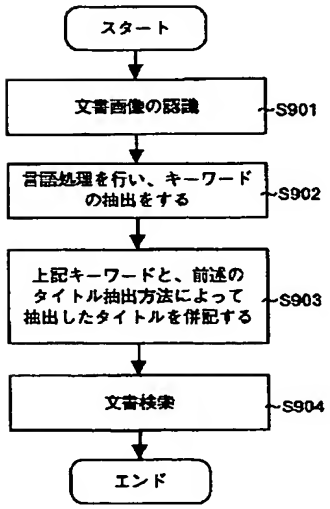
【図３】



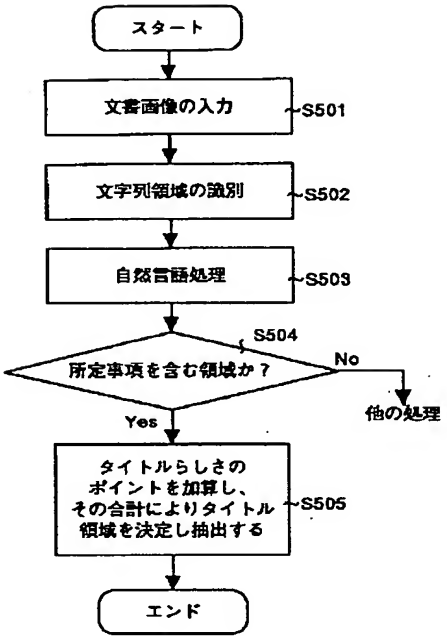
【図４】



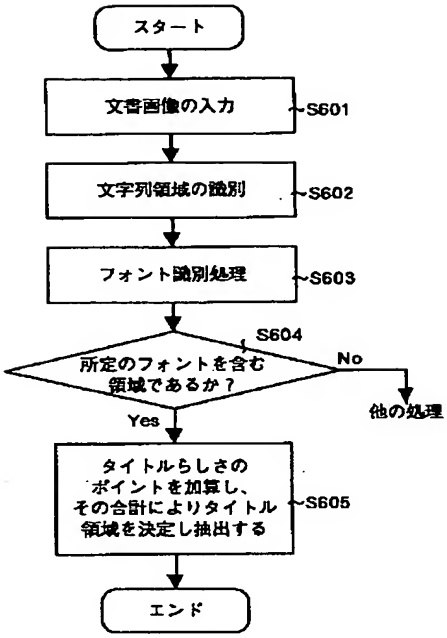
【図９】



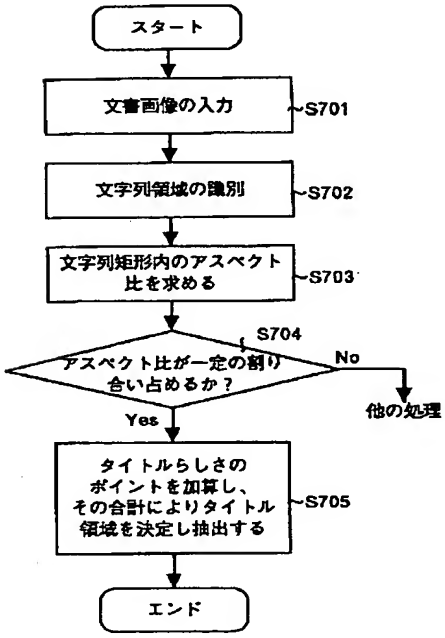
【図５】


スタート
文書画像の入力
S501
文字列領域の識別
S502
自然言語処理
S503
S504
所定事項を含む領域か？
No
他の処理
Yes
タイトルらしさの
ポイントを加算し、
その合計によりタイトル
領域を決定し抽出する
S505
エンド

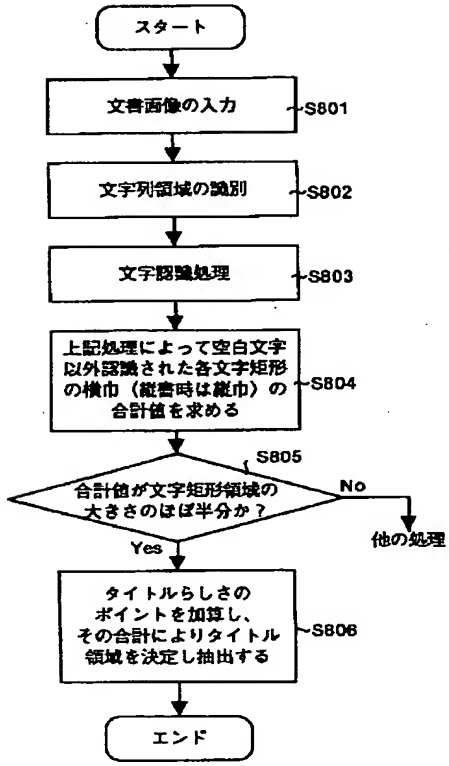
【図６】


スタート
文書画像の入力
S601
文字列領域の識別
S602
フォント識別処理
S603
S604
所定のフォントを含む
領域であるか？
No
他の処理
Yes
タイトルらしさの
ポイントを加算し、
その合計によりタイトル
領域を決定し抽出する
S605
エンド

【図７】


スタート
文書画像の入力
S701
文字列領域の識別
S702
文字列矩形内のアスペクト
比を求める
S703
S704
アスペクト比が一定の割り
合い占めるか？
No
他の処理
Yes
タイトルらしさの
ポイントを加算し、
その合計によりタイトル
領域を決定し抽出する
S705
エンド

【図８】


スタート
文書画像の入力
S801
文字列領域の識別
S802
文字認識処理
S803
上記処理によって空白文字
以外認識された各文字矩形
の横巾（縦書時は縦巾）の
合計値を求める
S804
S805
合計値が文字矩形領域の
大きさのほぼ半分か？
No
他の処理
Yes
タイトルらしさの
ポイントを加算し、
その合計によりタイトル
領域を決定し抽出する
S806
エンド